DOCUMENT RESUME

ED 257 859 TM 850 323

AUTHOR Yap, Kim Onn
TITLE Test Use Satisfaction: A Consumer Perspective.
INSTITUTION Northwest Regional Educational Lab., Portland, OR. Northwest Center for State Educational Policy Studies.
PUB DATE Mar 85
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Attitude Measures; *Educational Testing; Elementary Secondary Education; English (Second Language); Language Proficiency; Limited English Speaking; *Participant Satisfaction; Special Programs; *Teacher Attitudes; Test Construction; *Testing Problems; Test Use
IDENTIFIERS Hawaii (Honolulu); Students of Limited English Proficiency Program

ABSTRACT

This study investigated the relationship between a test's psychometric characteristics and school staff satisfaction with its use in programmatic activities. The Inventory of Test Use Satisfaction (IOTUS) was used to evaluate two individually administered oral language production tests used in Hawaii's Students of Limited English Proficiency (SLEP) Program: the Basic Inventory of Natural Language and the Language Assessment Scales. Administered to 142 SLEP Program Staff in three Honolulu school districts, the IOTUS surveyed attitudes about each test in four areas of test characteristics: measurement validity, examinee appropriateness, technical excellence, and administrative usability. The top five test characteristics which correlated significantly with test user satisfaction were reliability, item relevance and test bias, problems in test use construct validity, and experience with test administration. Results indicated only half of the respondants were satisfied with test use. Systematic and intensive involvement of test users in the test development process is suggested. (BS)

Test Use Satisfaction: A Consumer Perspective

Kim Onn Yap

Northwest Regional Educational Laboratory

A paper presented at the annual meeting

of the American Educational Research Association

Chicago, March 31 - April 4, 1985

TEST USE SATISFACTION: A CONSUMER PERSPECTIVE

The paper describes an attempt to identify test characteristics which appeal most to the average test user. A survey instrument, created on the basis of generally accepted test evaluation criteria and test use satisfaction, was administered to some 142 instructional staff in three local school districts. The results suggest that test users and psychometricians may not view the various test characteristics in the same perspective or attach the same importance to them. Implications of the findings for test construction and use are discussed in the paper.

Test Use Satisfaction: A Consumer Perspective

OBJECTIVE

A common set of criteria has generally been accepted by psychometricians for test evaluation. However, little, if any, research has been conducted to relate these test characteristics to test use satisfaction from the consumer's perspective. When over 200 million achievement tests are given annually in this country it is somewhat surprising that not much has been done to find out what really appeals to the average test user. The primary objective of this paper is to discuss the relationships between the psychometric qualities of a test and the degree to which school people are satisfied with its use in programmatic activities. It is hoped that the findings will help alert test developers to test characteristics having a direct bearing on test use satisfaction.

BACKGROUND

The Students of Limited English Proficiency (SLEP) Program in Hawaii serves students whose dominant language is not English and whose limitation in the use of English prevents them from functioning effectively in the regular classroom. The overall objective of the program is to help these students to adjust to the American culture in the Hawaiian setting by acquiring basic communication skills to

participate in the regular classroom instruction and school activities appropriate for their age and grade level.

Students are selected to participate in the program on the basis of their language dominance ratings as determined by criteria specified in the Identification Assessment Programming System (Hawaii Department of Education, 1980). Only students who receive language dominance ratings of 1 and 2 are eligible to participate in the program. Participants are exited from the program when they reach a language dominance rating of 3 (or above) and score at the 25 percentile (or above) on a standardized test in reading, language arts, and mathematics.

The SLEP Program is offered in all seven school districts in the state. These districts use either the Basic Inventory of Natural Language (BINL) or the Language Assessment Scales (LAS) to measure English language proficiency of participating students. Both the BINL and the LAS are individually administered oral language production tests.

In 1983 a study was conducted to review and evaluate the respective tests. In the course of the study, surveys were conducted in the Honolulu, Leeward and Central districts to obtain a measure of test use satisfaction from the respective project staff. The surveys examined the relationships between the psychometric qualities of the tests and the degree to which school people are satisfied with their use in programmatic activities.

PROCEDURE

Criteria

Several sources were used to develop a set of criteria for test evaluation. These included documents produced by the Center for the Study of Evaluation of UCLA (Hoepfner, et al., 1976), the Center for Bilingual Education (Silverman, et al., 1976; Silverman, et al., 1978) and the Assessment Projects at the Northwest Regional Educational Laboratory (Nafziger, et al., 1975), the American Psychological Association, the American Educational Research Association, the National Council of Measurement in Education (Davis, 1974), as well as individual researchers (e.g., Madaus, et al., 1982). The final set of criteria used in the present study thus represents a comprehensive compilation of generally accepted test standards which had been field tested and used in test evaluation.

More specifically, the criteria relate to four major areas of test characteristics: measurement validity, examinee appropriateness, technical excellence, and administrative usability. The criterial areas are further described as follows:

Measurement validity. This set of criteria looks at the nature of what a test measures, the range of behaviors sampled, the relationship of the test score to other measures, and the demonstrated usefulness of the test in theoretical or practical settings.

Examinee appropriateness. These criteria relate to the appropriateness of the test materials including content of the stimuli (items) and mode of response, relative to the grade level of students taking the test.

Administrative usability. These criteria deal with practical concerns in administering and using a test. The ease with which the test can be given, scored, and interpreted, and the usefulness of the resulting score in making program or instructional decisions.

Technical excellence. These criteria are concerned with the test's reliability, replicability and refinement of measurement.

Instrument

A survey instrument, the Inventory of Test Use Satisfaction (IOTUS), was developed on the basis of the identified criteria for test evaluation. The instrument consisted of two parts. Part I was made up of 6 items relating to the respondent's general knowledge of and experience with the test in question. Part II consisted of 35 items relating to the specific test evaluation criteria and test use satisfaction on a five-point Likert scale (SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree). Items specific to information contained in the BINL or LAS manual were excluded. A sample of survey items follows:

- I know what the test is supposed to measure.
- The test measures something distinct from what is measured by other similar tests.
- The test provides reliable information for its intended use.
- The test items are relevant to my students.
- I am satisfied with the use of the test in my program.

Test Use Survey

Three separate surveys were conducted in the Honolulu (LEA 1), Leeward (LEA 2) and Central (LEA 3) districts in April 1983 to obtain a measure of test use satisfaction on the BINL and the LAS. Data obtained from the surveys were coded and entered into the computer by district staff. Analyses were performed separately for each district.

Frequencies of responses were tabulated and corresponding percentages were calculated. To convert the Likert-scale data to dichotomous (i.e., yes-no) data, adjacent response categories (e.g., SD and D) were combined. Neutral responses were treated as a separate category. Inter-item correlations were computed for all IOTUS items, separately for each district. Correlations between items pertaining to test characteristics and test use satisfaction were then singled out for examination.

RESULTS

A total of 142 SLEP Program staff responded to the surveys, providing a response rate of over 73 percent. The respondents included teachers, part-time temporary teachers and educational assistants. A predominant majority of the respondents (over 80 percent) rated their knowledge of the respective tests as good or excellent. Over 90 percent reported that they had administered the respective tests 8 or more times. Slightly more than one-half (53 percent) of the respondents indicated that they

were satisfied with test use.

Twelve test characteristics were found to correlate significantly ($p < .05$) with test use satisfaction for all three districts. These items pertained to:

- content coverage (.39 - .80)
- conceptual soundness of items (.41 - .50)
- construct validity (.60 - .69)
- concurrent validity (.28 - .63)
- reliability (.50 - .76)
- quality of experience in test administration (.45 - .77)
- information for program improvement (.48 - .64)
- information for instructional decisions (.42 - .65)
- item relevance (.60 - .69)
- problems in test use (.48 - .72)
- range of raw scores (.46 - .70)
- range of converted scores (.40 - .63)

When these items were ranked on the magnitude of their correlations with test use satisfaction, the top five test characteristics included:

- reliability
- item relevance
- problems in test use
- construct validity
- experience with test administration

As shown in Table 1, while there were inter-district differences, these correlations ranged from moderate (.48) to quite substantial (.77) in size. The importance of reliability and construct validity is

apparent to psychometricians. Item relevance, problems in test use and experience with test administration reflect a distinct user orientation.

==========
Table 1 about here
==========

With regard to reliability, about two-thirds (68 percent) of the respondents indicated that the respective tests provided reliable information for its intended use. Interestingly enough, a higher percentage (75 percent) of the project staff reported using test information for various purposes (e.g., evaluation and student selection).

None of the tests included in the study apparently posed any serious problems in terms of test administration. A great majority of the respondents (89 percent) reported positive experience with test administration, indicating that (a) they had no difficulty in administering the respective tests; (b) they were able to administer the tests in the same way each time they tested their students; and (c) the way in which students were required to respond to test items was simple and direct. These respondents further indicated that the respective test manuals were clear, well-organized, consistent, thorough and helpful.

The respondents told quite a different story with respect to construct validity and item relevance. Fewer than one-half (41 percent) of them believed that the respective tests measured something distinct from what was measured by other similar tests. Even fewer (26 percent) believed that the respective tests provided results capable of predicting

how well students may perform in other school subjects. Only a small proportion (36 percent) of the respondents reported that the items in the tests were relevant to their students. Fewer than one-half (48 percent) indicated that the test items were free of cultural, sexual and ethnic bias.

DISCUSSION

The level of test use satisfaction revealed in the present study must be described as low. Approximately one-half (53 percent) of the project staff surveyed reported that they were satisfied with test use. Several test characteristics have apparently contributed to the low level of test use satisfaction. The most notable are discussed below.

First, while both the BINL and the LAS have reportedly satisfactory reliability (Herbert, 1979; De Avila and Duncan, 1977), only two-thirds of the project staff indicated that their test provided reliable information for its intended use. Perhaps more interestingly, a higher percentage (75 percent) of the respondents reported using the test information for various purposes, suggesting that some project staff used the test data even when they felt the data might not have sufficient reliability.

Secondly, a much smaller proportion of the respondents (60 percent) used the test data for instructional purposes (e.g., diagnosis and instructional planning). With the current thrust of educational reform moving toward program improvement, teachers and school administrators are eagerly seeking information for instructional improvement. Failure to provide such information represents a serious drawback of many existing standardized achievement tests. Test publishers not only should make such information available, but also provide ways of using such

information for program improvement (e.g., Wilson and Hiscox, 1984).

Thirdly, construct validity has long been a difficult test characteristic to measure and to demonstrate. Not only should a test measure some mental or behavioral entity, it also should measure some entity not measured by similar tests. Its existence is difficult to justify, otherwise. In the field of oral language measurement, many tests have been developed for use with bilingual students (Silverman, et al., 1976). If the findings in the present study are any indication, many of these tests probably measure the same things, if they measure anything at all. Only 41 percent of the respondents believed that their respective tests measured something distinct from what was measured by other similar tests. This clearly suggests that test developers still have a long way to go in their efforts to adequately assess language proficiency of special target groups.

Fourthly, no other problems are more dramatically highlighted than the problem of item relevance and test bias in the present study. Slightly more than one-third (36 percent) of the respondents indicated that the items in their respective tests were relevant to their students. Moreover, fewer than one-half (48 percent) of them reported that the items were free of cultural, sexual and ethnic bias. It would seem that after years of research (Subkoviak et al., 1984; Van der Flier et al., 1984) item relevance and test bias remain an elusive and perennial problem for measurement experts and test developers. While more research is obviously needed, systematic and intensive involvement of test users in the test development process would perhaps be more conducive to solving the problem. Investment by test publishers in this aspect of test development promises considerable payoff in terms of test use satisfaction.

REFERENCES

Davis, F. B. (1974). Standards for Educational and Psychological Tests. Washington, D.C.: American Psychological Association.

De Avila, E. A., and Duncan, S. E. (1977). Language Assessment Scales. Corte Madera, CA: Linquametrics Group.

Hawaii State Department of Education (1980). Identification, Assessment and Programming System for Students of Limited English Proficiency: A Systems Manual 1980-81. Honolulu: Hawaii State Department of Education.

Herbert, C. H. (1979). Basic Inventory of Natural Language. San Bernardino, CA: CHECpoint Systems.

Hoepfner, R., Bastone, M., Ogilvie, V., Hunter, R., Sparta, S., Grothe, C. R., Shani, E., Hufano, L., Goldstein, E., Williams, R, and Smith, K. O. (1976). CSE Elementary School Test Evaluations. Los Angeles: Center for the Study of Evaluation, UCLA.

Madaus, G. G., Airasian, P. W., Hambleton, R. K., Consalvo, R. W., and Orlandi, L. R. (1982). Development and application of criteria for screening commercial, standardized tests. Educational Evaluation and Policy Analysis, 4(3), 401-415.

Nafziger, D. A., Thompson, R. B., Hiscox, M. D., and Owen, T. R. (1975). Tests of Functional Adult Literacy: An Evaluation of Currently Available Instruments. Portland, OR: Northwest Regional Educational Laboratory.

Silverman, R., Noa, J. K., and Russell, R. H. (1976). Oral Language Tests for Bilingual Students: An Evaluation of Language Dominance and Proficiency Instruments. Portland, OR: Northwest Regional Educational Laboratory.

Silverman, R., and Tupper, N. (1978). Assessment Instruments in Bilingual Education. Portland, OR: Northwest Regional Educational Laboratory.

Subkoviak, J. J., Mack, J. S., Iranson, G. H., and Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21(1), 49-58.

Van der Flier, H., Mellenbergh, G. J., Ader, H. J., and Wijn, M. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21(2), 131-145.

Wilson, S. M., and Hiscox, M. D. (1984). Using standardized tests for Assessing local learning objectives. Educational Measurement Issues and Practice, 3(3), 19-22.

Table 1

Relationships Between Test Characteristics

and Test Use Satisfaction

Test Characteristics	Correlation Coefficients		
	LEA 1 (N=61)	LEA 2 (N=45)	LEA 3 (N=36)
Reliability	.50	.76	.76
Item relevance	.65	.60	.69
Problems in test use	.48	.62	.72
Construct validity	.69	.61	.60
Experience with test administration	.52	.45	.77